Title of Project: Semi-Automated Identification of Biomedical Literature

Principal Investigator: Thomas Trikalinos

Team Members: Haris Papageorgiou, Dimitris Pappas, Evangelos Evangelou, Gaelen P. Adam

Project Period Begin Date: 09/30/2019

Project Period End Date: 09/30/2020

Federal Project Officer: Lionel Banez, MD

Acknowledgment of Agency Support: The work described here was supported by grant No R03 HS027247-01 from the Agency for Healthcare and Quality (AHRQ) Brown University.

Grant Award Number: R03 HS027247

Table of Contents:

Structured Abstract

Purpose: The main objective of the project was to design and develop a state-of-the-art information system assisting experts in conducting literature discovery for evidence-synthesis products in an efficient and effective way. The system unifies the query formulation and citation screening processes by exploiting modern machine learning and information retrieval methods.

Scope: The typical approach to literature identification involves two discrete and successive steps: (i) formulating a search strategy (i.e., a set of Boolean queries) and (ii) manually identifying the relevant citations in the corpus returned by the query. We have developed a literature identification process that unifies the query formulation and citation screening steps and uses modern approaches for text encoding (dense text embeddings) to represent the text of the citations in a form that can be used by information retrieval and machine learning algorithms.

Methods: The system described in this report takes as input a set of questions written in natural language, generated from the report key questions and uses a set of machine-learning algorithms to rank all of PubMed's citations based on relevance to each question. It then returns the top-ranked citations for each question to a total of 100 citations. The first 100 articles are exported and screened manually, with the manual screener adjudicating the relevance of each abstract and tagging words that indicate relevance in relevant abstracts. These "curated" articles are then exploited by the system to refine the search and re-rank the abstracts, and a new set of 100 relevant abstracts is exported and screened/tagged until convergence (i.e., no other relevant abstracts could be retrieved) or for a certain number of iterations (batches), which we set to10 in our experiments. System performance was assessed, using seven ongoing or completed systematic reviews (three prospectively and four retrospectively). Sensitivity, precision, and the number of abstracts needed to read were calculated for each project.

Results: The ability of the system to identify the relevant articles varied across reviews from a low of 0.16 for Sleep Apnea to a high of 0.58 for Diverticulitis. HIP had a lower sensitivity (0.08), but this was across only 6 batches. For nearly all reviews, the precision was drastically improved compared to the standard procedure of separate searching and manual screening, ranging from 0.01 to 0.09 (NNR 87 to 11) as compared to 0.006 to 0.083 (NNR 143 to 12) for the standard two-step process. Looking at factors that might affect sensitivity, we found that generally the reviews that had greater overall sensitivity retrieved more relevant citations in early batches, but neither study design, study size, nor specific key question significantly affected retrieval across reviews.

Discussion: Future research should explore ways to encode domain knowledge in query formulation, possibly by incorporating a "reasoning" aspect to the system to elicit more contextual information and leveraging ontologies and knowledge bases to better enrich the questions used in the search.

Key Words: evidence synthesis, systematic review methods, literature identification, abstract screening, text mining, machine learning

Purpose

The main objective of the project was to design and develop a state-of-the-art information system to assist experts in conducting literature discovery for systematic reviews, scoping reviews, rapid reviews, and related evidence-synthesis products, which are key tools in Evidence-based Medicine (EBM) and patient-centered comparative effectiveness research. The need to modernize the methods and processes of EBM is pressing, in part because of the rate with which new information is generated.[1-3] For example, in 2010, Bastian et al. estimated that 75 new trials and 11 systematic reviews were entered in PubMed every day[4]; these numbers grew to approximately 100 trials and 75 systematic reviews in 2017 (own data). This suggests that, over time, an increasing number of systematic reviews and related products are being completed, and that each of them will have to examine larger evidence-bases and need more frequent updating.

Scope

The typical approach to literature identification involves two discrete and successive steps: (i) formulating a search strategy (normally, a set of Boolean queries) and (ii) manually screening the relevant citations in the corpus returned by the query/ies, preferably by at least 2 people.[5] This ensures that the literature identification process is transparent and replicable, but not necessarily that it is comprehensive. A systematic review of empirical methodological studies on literature identification suggests that a substantial number of relevant citations are missed in searches conducted by expert librarians.[6] This is why it is recommended to search multiple databases and registries that overlap in scope, and to supplement database queries by perusing the reference lists of relevant articles.[5] We have developed a literature identification process that substantially departs from the above paradigm in that it (i) unifies the query formulation and citation screening steps and (ii) uses modern deep learning and information retrieval methods to increase the efficiency and effectiveness of literature identification. In this system, we use modern approaches for semantic text encoding (dense text embeddings) to represent the text of the citations in a form that can be used by deep neural networks (which is the state of science today in text classification and ranking).

Methods

In the traditional method, a human (usually a trained medical librarian) creates a search strategy based on the review's population and intervention (or exposure), along with possible other concepts that may include outcomes, study designs, language, location, and so on. The librarian identifies a comprehensive set of synonyms and controlled vocabulary terms for each concept of interest, then combines them into one or more queries, using Boolean logic. These queries are then manually executed in each database. Subsequently, members of the review team double-screen each citation returned by the query strategy.

In contrast, the system described in this report takes as input a set of questions written in natural language, generated from the report key questions, and uses a set of machine-learning algorithms to rank all of PubMed's citations based on relevance to each question. It then returns the top-ranked citations for each question to a total of 100 citations (e.g., if there are two questions, it selects the top 50 for each question; if there are 10 questions, it selects only the top ten for each question). A citation can appear in the results for more than one question. Based on the screening decisions and the annotation of relevant terms in previous batches, the system refines its search and returns the next top 100 unscreened citations.

Creating the Dataset

The literature collection was constructed using the metadata of biomedical articles accessible through PubMed (ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline/). For each article, we concatenated the title and abstract of each citation and indexed the resulting concatenated text in an ElasticSearch engine (https://www.elastic.co/elasticsearch/), along with the publication date of the paper. We discarded any document without an abstract, ending up with approximately 21 million articles from the original 31 million articles.

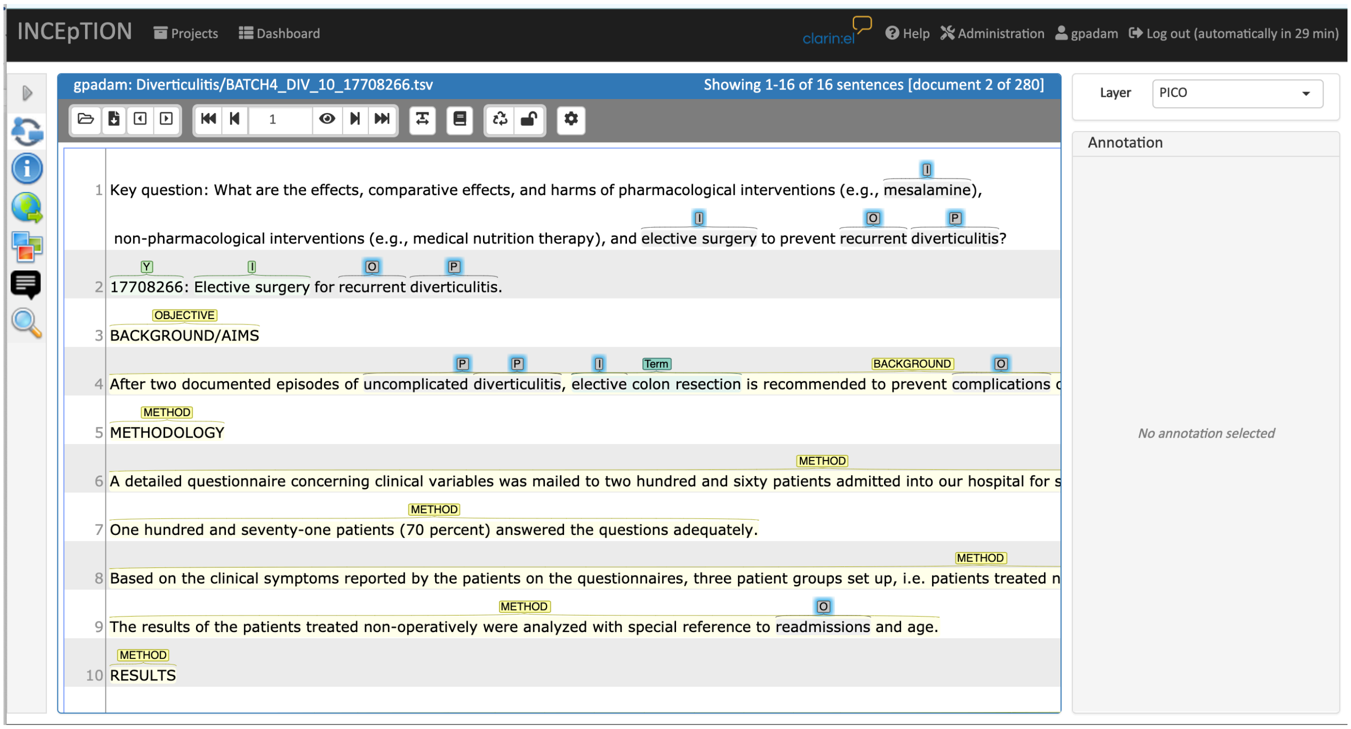
Selecting the Citations for Screening

We developed an information system which is fed with experts' questions, formulated by in natural language, along with any articles pre-annotated as relevant or irrelevant. The system extracts a set of positive key-phrases from the documents that have already been annotated as positive and a set of negative key-phrases from the negatively annotated documents, using the unsupervised key-phrase extraction algorithm SGRank (https://www.aclweb.org/anthology/S15-1013/). Each expressed question is then executed, and 200 articles are retrieved using theBM25 retrieval algorithm, excluding any abstracts that had been previously retrieved. We then re-ranked the retrieved articles using a deep-learning model for joint document and snippet retrieval (http://nlp.cs.aueb.gr/pubs/aueb_at_bioasq7.pdf). We penalized the score of each article containing a negative key-phrase and increased the score of any article containing a positive key-phrase, thus taking into account the feedback provided by users. Finally, the system returned the top 100 documents for human evaluation.

Experimental setup

The first 100 articles were exported and screened manually, with the manual screener adjudicating the relevance of each abstract and tagging words that indicate relevance by PICO element in those abstracts that were deemed relevant (See example abstract in Figure 1). These "curated" articles were used by the system to refine the search and re-rank the abstracts, and a new set of 100 relevant abstracts was exported to be screened/tagged. For this project, we chose to limit each review to 10 batches, representing between 800 and 1000 unique citations screened. Due to time limitations, we were only able to do 6 batches (536 unique citations) for the Headaches in Pregnancy project.

Figure 1. Example of an annotated abstract



gpadam: Diverticulitis/BATCH4_DIV_10_17708266.tsv

Showing 1-16 of 16 sentences [document 2 of 280]

1 Key question: What are the effects, comparative effects, and harms of pharmacological interventions (e.g., mesalamine), non-pharmacological interventions (e.g., medical nutrition therapy), and elective surgery to prevent recurrent diverticulitis?

2 17708266: Elective surgery for recurrent diverticulitis.

3 BACKGROUND/AIMS

4 After two documented episodes of uncomplicated diverticulitis, elective colon resection is recommended to prevent complications c

5 METHODOLOGY

6 A detailed questionnaire concerning clinical variables was mailed to two hundred and sixty patients admitted into our hospital for s

7 One hundred and seventy-one patients (70 percent) answered the questions adequately.

8 Based on the clinical symptoms reported by the patients on the questionnaires, three patient groups set up, i.e. patients treated n

9 The results of the patients treated non-operatively were analyzed with special reference to readmissions and age.

10 RESULTS

Performance measures

The retrieved citations were divided into different subgroups, as per Table 1. The top row of the table includes all citations identified by the system and screened (P), divided by whether they were included in the final report (TP) or not (FP). The bottom row includes all citations not identified by the system. Because this number is extremely large (our source set included upwards of 21 million citations), we are only interested in the citations included in the report that were not identified by the system (FN). I includes all citations in the report that have a PubMed identifier (PMID).

Table 1. Cross-classification matrix for semi-automated screening.

	Citations included in report	Citations not included in report	Total
Identified by the system	TP	FP	P
Missed by the system	FN		
Total	I		

TP: Identified by the system and ultimately included in the systematic review. FN: missed by the system, but ultimately included in the review. FP: System predicted relevant but ultimately excluded from the review. P: all citations identified by the system. I: all citations eventually included in review with a PubMed identifier.

The aim of citation screening with active learning is to minimize the workload while being as accurate as possible. Thus, we were interested in two dimensions of classification performance: workload and sensitivity (i.e., recall). Our aim was to maximize sensitivity while minimizing workload (operationalized here as precision and its inverse, number needed to read [NNR]).

Sensitivity is the ability to identify the truly relevant citations in the computer-assisted screening process:

$$\text{Sensitivity} = TP/I$$

To measure the workload, we define precision as the proportion of citations screened that were relevant:

$$\text{Precision} = TP/P$$

To make this number more intuitive, we use Number Needed to Read (NNR), defined as the number of irrelevant citations that the reviewer had to screen for each relevant citation found:

$$\text{NNR} = 1/\text{Precision}$$

For comparison, we report precision and NNR for the manual screening process as defined by the number of relevant articles included in the final report as a proportion of the total number of articles retrieved in the PubMed searches for each review.

Results

We selected seven recently completed systematic reviews (details in Table 2). In three of these reviews, the system was used prospectively, with a human annotator screening ten batches of 100 citations each. In the other four reviews, the system was evaluated retrospectively, with automatic annotation. The level 1 (human abstract screening from the original review) labels were used for this evaluation. The human feedback was simulated using a list of known positive and negative articles. In the retrospective evaluation we examined three settings, namely "retro res", "retro res no KTs" and "retro res no neg KTs". In the retro res no KTs setting, we did not use the key-phrases of the documents. In the retro res no neg KTs setting, we used only the positive key phrases to increase the score of all articles that share a key phrase with a known positive article. Finally, in the retro-res setting, we penalized any document that share a key phrase with a known negative article and increased the score of any document that shares a key phrase with a known positive article. In all retrospective settings, we may have missed any relevant article that was not included in the provided list (i.e., reviewers of the original review never saw that article). Therefore, we expect that retrospective scores would improve with human inspection.

Table 2. Included Datasets

Prospective/ retrospective	Systematic Review Dataset (reference)	Brief description	Domain	Citations screened in full review, N (N from PubMed)	Screened in at title/ abstract/ keyword level (% of n)	Screened in based on full text (% of n); N (%) from PubMed	No. KQs	No. NLQs
Prospective	**Diverticulitis:** Management of Colonic Diverticulitis[7]	Evaluates the effectiveness and harms of various nonsurgical treatment options for acute diverticulitis, clinical consequences of diagnostic imaging, detection strategies for colorectal cancer (CRC) in patients with recent diverticulitis, and preventive options for long-term recurrence.	Surgical	15,199 (7,981)	722 (4.75)	88 (0.6); 86 (1.1)	4	14
Prospective	**Sleep Apnea (SA):** Continuous Positive Airway Pressure Treatment for Obstructive Sleep Apnea[8]	Summarizes evidence on long-term clinical health outcomes with CPAP treatment and assesses the validity of surrogate and intermediate measures (e.g., AHI) for clinically significant outcomes.	Sleep Medicine	15,333 (10,891)	1,593 (10.4)	71 (0.5); 70 (0.6)	2	7
Prospective	**Headaches in Pregnancy (HIP):** Management of Primary Headaches in Pregnancy[9]	Evaluates the literature on pharmacologic and nonpharmacologic interventions to prevent or treat attacks of primary headaches in women who are pregnant, attempting to become pregnant, postpartum, or breastfeeding.	Obstetrics	8,154 (6,587)	400 (4.9)	72 (0.8); 64 (9.7)	2	6
Retro-spective	**Nonmelanoma Skin Cancer (NMSC):** Treatments for Basal Cell and Squamous Cell Carcinoma of the Skin[10]	Comprehensively collects information on the comparative effectiveness and safety of each currently used therapeutic strategy for both BCC and SCC	Oncology	15,813 (9,741)	534 (3.4)	125 (0.8); 78 (0.8)	2	2

Prospective/ retrospective	Systematic Review Dataset (reference)	Brief description	Domain	Citations screened in full review, N (N from PubMed)	Screened in at title/ abstract/ keyword level (% of n)	Screened in based on full text (% of n); N (%) from PubMed	No. KQs	No. NLQs
Retro-spective	**Tympanostomy Tubes (Tymp.):** Tympanostomy Tubes in Children With Otitis Media[11]	Synthesize information on the effectiveness of tympanostomy tubes (TT) in children with chronic otitis media with effusion and recurrent acute otitis media, summarize the frequency of adverse effects or complications associated with TT placement, synthesize information on the necessity for water precautions in children with TT, and assess the effectiveness of available treatments for otorrhea in children who have TT	Pediatrics	8,498**	509(6.0)**	175 (2.0)**	5	8
Retro-spective	**Urinary Incontinence (UI):** Nonsurgical Treatments for Urinary Incontinence in Women: A Systematic Review Update[12]	Systematically review and meta-analyze the comparative effectiveness and harms of nonpharmacological and pharmacological interventions for women with all forms of UI.	Urology	7,840 (3,706)	723 (9.2)	244 (3.1); 96 (2.6)	4	24
Retro-spective	**Venous Thromboembolism (VTE):** Venous Thromboembolism Prophylaxis in Major Orthopedic Surgery: Systematic Review Update	Systematically review the comparative effectiveness for VTE outcomes and harms of different thromboprophylaxis interventions for patients undergoing major orthopedic surgery (THR, TKR, and HFx surgery).	Surgical	1,738 (642)	455 (26.2)	56 (3.2); 54 (8.4)	6	8

KQ=key question; NLQ=natural language question. Numbers may vary slightly from those given in the final report due to the stage of the report at the time of this analysis.
** The PubMed search was done separately for this review, so only PubMed numbers are given.

Sensitivity and Precision/NNR

Table 3 gives the overall sensitivity, precision, and number needed to read (NNR) for each review. The ability of the system to identify the relevant articles varied across reviews from a low of 0.16 for Sleep Apnea to a high of 0.58 for Diverticulitis. HIP had a lower sensitivity (0.08), but this was across only six batches. For nearly all reviews, the precision was drastically improved compared to the standard procedure of separate searching and manual screening, though again the actual precision varied across reviews, ranging from 0.01 to 0.09 (NNR 107 to 11). By comparison, the precision of the original search ranged from 0.006 to 0.097 (NNR 167 to 10).

Table 3. Sensitivity and precision/NNR.

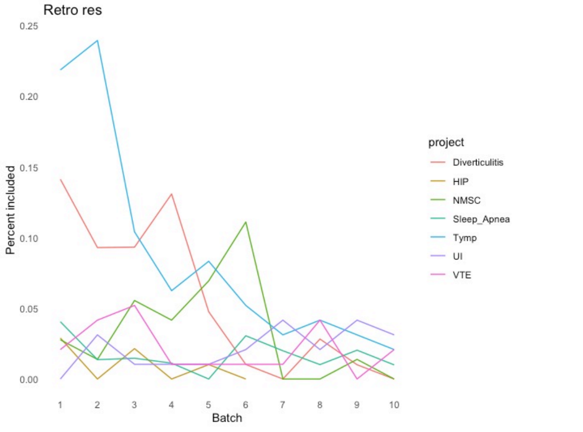
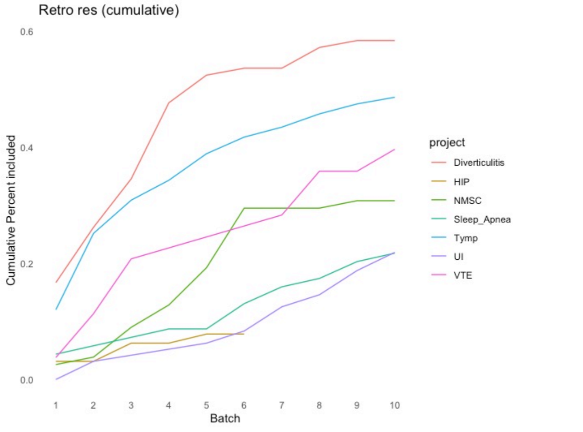
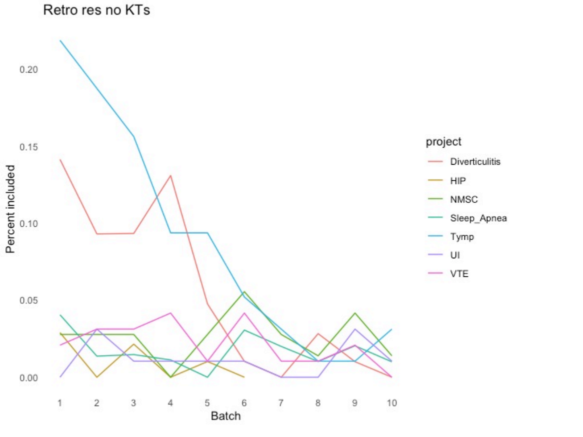
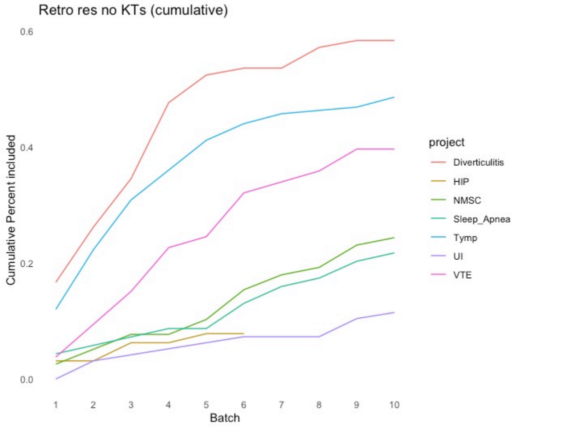
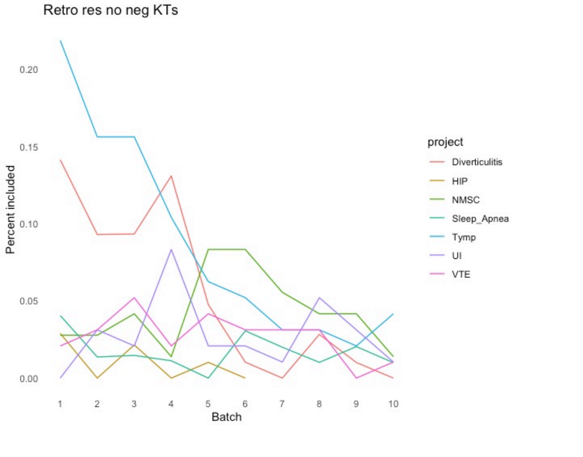
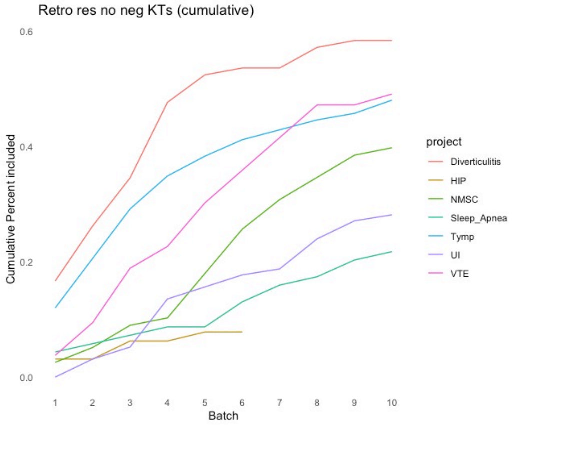
Review	Iteration	Sensitivity	Precision	NNR	Precision for entire search	NNR for entire search
Prospective						
Diverticulitis	Overall	0.58 (49/84)	0.05	19	0.011*	91
HIP**	Overall	0.08 (5/64)	0.01	107	0.097*	10
Sleep Apnea	Overall	0.16 (11/69)	0.01	80	0.006*	167
	Part 1	0.07 (5/69)	0.01	80		
	Part 2	0.10 (7/69)	0.01	72		
Retrospective						
NMSC	Retro res	0.31 (24/78)	0.03	30	0.008*	125
	Retro res no KTs	0.24 (19/78)	0.03	38		
	Retro res no neg KTs	0.40 (31/78)	0.04	23		
Tymp.	Retro res	0.48 (85/175)	0.09	11	0.020*	49
	Retro res no KTs	0.48 (85/175)	0.09	11		
	Retro res no neg KTs	0.48 (84/175)	0.09	11		
UI	Retro res	0.22 (21/96)	0.02	46	0.026*	38
	Retro res no KTs	0.11 (11/96)	0.01	87		
	Retro res no neg KTs	0.28 (27/96)	0.03	36		
VTE	Retro res	0.40 (21/53)	0.02	46	0.083*	12
	Retro res no KTs	0.38 (20/53)	0.02	48		
	Retro res no neg KTs	0.47 (25/53)	0.03	38		

NNR = number needed to read; HIP = headaches in pregnancy; NMSC = nonmelanoma skin cancer; Tymp. = tympanostomy tubes; UI = urinary incontinence; VTE = venous thromboembolism.
*includes citations from PubMed only; some of the citations with PMIDs may have been found through other databases, but most were probably identified in the PubMed search.
** For headaches in pregnancy, we were only able to run 6 batches, as compared to 10 for all other projects.

Sensitivity and Precision by iteration (batch)

Figure 2 shows the proportion of relevant articles identified by batch. The left-hand column shows the proportion of relevant articles identified in each batch (relevant identified/total screened), while the right-hand column shows the cumulative percentage of relevant articles found across batches (cumulative relevant identified/total relevant). There is a similar pattern across iterations of the retrospective reviews, and generally the figures show that the reviews with greater overall sensitivity retrieved more relevant citations in early batches.

Figure 2. Sensitivity by batch.

Analysis	By batch	Cumulative
Retro res		
Retro res no KTs		
Retro res no neg KTs		

For headaches in pregnancy, we were only able to run 6 batches, as compared to 10 for all other projects.

Factors affecting sensitivity

Tables 4 through 8 examine the results for each review, in terms of specific aspects that may have made a given citation systematically more likely to be identified. In general, neither study design, study size, nor specific key question significantly affected retrieval across reviews.

There is some indication that the algorithm is more likely to find randomized controlled trials (RCTs) than other study designs, as the percentage of RCTs found is consistently higher across datasets, but for only one dataset (Tympanostomy) was this difference statistically significant. For two datasets, we found a statistically significant difference across key questions. For Diverticulitis, this was driven by a large percentage of the Key Question 2 (antibiotic treatment 76%) and 3 (colonoscopy 84%) studies (Table 4); for Tympanostomy (Table 5), it was driven by the retrieval of a very high percentage of the Key Question 5 (treatment of otorrhea 93%) studies. A third dataset, Non-melanoma Skin Cancer, showed a nearly significantly larger percentage of Key Question 2 (adverse event) studies in the Retro res iteration, but in this review Key Question 2 comprised a subset of Key Question 1 (effectiveness) studies (Table 7). Finally, in the Venous Thromboembolism (VTE) dataset, there was a significant difference among key questions for two of the three tests, and a near significant difference in the third, with a large percentage of the key question 5 studies (those included in the network meta-analysis of efficacy) identified and small percentages of the key question 4 (combined versus single interventions) and 6 (timing of VTE prophylaxis) studies identified (Table 8).

Table 4. Subgroup analysis for prospective datasets

	Diverticulitis			*Sleep Apnea*			*Headaches in Pregnancy*		
		found/total (%)	**P-value**		**found/total (%)**	**P**		**found/total (%)**	**P-value**
Study design	RCT	18/28 (64)	0.5390	RCT	8/35 (23)	0.3682	RCT	0/2 (0)	0.0778
	NRCS	13/22 (59)		NRCS	2/21 (10)		NRCS	0/13 (0)	
	cohort/ single arm	16/34 (47)		cohort/ single arm	1/13 (8)		cohort/ single arm	2/5 (40)	
							Case report	2/18 (11)	
							Systematic Review	1/27 (4)	
Key Question	KQ 1	3/7 (43)	***< 0.001***	KQ 1	9/65 (14)	0.5426	KQ 1	1/14 (7)	1.0000
	KQ 2	19/25 (76)		KQ 2	6/29 (21)		KQ 2	4/51 (8)	
	KQ 3	16/19 (84)							
	KQ 4	12/50 (24)							
Study size	<100	4/6 (67)	0.3674	<100	3/22 (14)	0.9468	1	2/18 (11)	0.6108
	101-500	26/42 (62)		101-500	6/31 (19)		2-200	2/8 (25)	
	501-1000	10/16 (63)		501-1000	0/3 (0)		201-1,000	0/4 (0)	
	1001-10000	7/20 (35)		>1001	2/13 (8)		1,000-5,547	0/7 (0)	

* Note: this does not include data that was not reported, and some studies fit into more than one category. P-value based on Fisher exact test.

Table 5. Subgroup analysis for retrospective datasets: Tympanostomy

		Retro res		*Retro res no KTs*		*Retro res no neg KTs*	
		found/total (%)	**P-value**	**found/total (%)**	**P-value**	**found/total (%)**	**P-value**
Study design	RCT	39/59 (66)	***0.0034***	39/59 (66)	***0.0034***	38/59 (66)	***0.0034***
	NRCS	10/33 (30)		10/33 (30)		10/33 (30)	
	cohort/single arm	33/68 (49)		33/68 (49)		33/68 (49)	
Key Question	KQ 1	17/54 (31)	***0.0009***	17/54 (31)	***0.0009***	17/54 (31)	***0.0009***
	KQ 2	6/13 (46)		6/13 (46)		6/13 (46)	
	KQ 3	41/90 (46)		40/90 (44)		39/90 (43)	
	KQ 4	6/11 (55)		6/11 (55)		6/11 (55)	
	KQ 5	13/14 (93)		13/14 (93)		13/14 (93)	
Study size	14-78	4/34 (41)	0.3210	13/34 (38)	0.0606	14/34 (41)	0.0995
	79-185	18/33 (55)		16/33 (48)		15/33 (46)	
	186-358	21/35 (60)		22/35 (63)		21/35 (60)	
	359-217,206	21/34 (62)		23/34 (68)		23/34 (68)	

* Note: this does not include data that was not reported, and some studies fit into more than one category. P-value based on Fisher exact test.

Table 6. Subgroup analysis for retrospective datasets: Urinary Incontinence

		Retro res		*Retro res no KTs*		*Retro res no neg KTs*	
		found/total (%)	**P-value**	**found/total (%)**	**P-value**	**found/total (%)**	**P-value**
Study design	RCT	17/80 (21)	0.4813	10/80 (13)	1.000	23/80 (29)	1.000
	NRCS	0/1 (0)		0/1 (0)		0/1 (0)	
	cohort/single arm	5/15 (33)		1/15 (7)		4/15 (27)	
Key Question	KQ 1	17/68 (25)	0.7244	8/68 (12)	0.6323	24/68 (35)	0.0663
	KQ 2	3/23 (13)		2/23 (9)		2/23 (9)	
	KQ 3	1/4 (25)		1/4 (25)		1/4 (25)	
	KQ 4	1/5 (20)		0/5 (0)		1/5 (20)	
Study size	12-42	3/25 (12)	0.1567	2/25 (8)	0.8154	3/25 (16)	0.8154
	43-90	8/22 (36)		2/22 (9)		11/22 (50)	
	91-184	3/24 (13)		4/24 (17)		7/24 (29)	
	184-12,733	6/25 (24)		3/25 (12)		5/25 (20)	

* Note: this does not include data that was not reported, and some studies fit into more than one category. P-value based on Fisher exact test.

Table 7. Subgroup analysis for retrospective datasets: Nonmelanoma Skin Cancer

		Retro res		*Retro res no KTs*		*Retro res no neg KTs*	
		found/total (%)	**P-value**	**found/total (%)**	**P-value**	**found/total (%)**	**P-value**
Study design	RCT	21/62 (34)	0.3642	16/62 (26)	0.7477	27/62 (44)	0.2537
	NRCS	3/16 (19)		3/16 (19)		4/16 (25)	
Key Question	KQ 1	24/78 (31)	0.0589	19/78 (24)	0.1431	31/78 (40)	0.2919
	KQ 2	10/18 (56)		8/18 (44)		10/18 (56)	
Study size	12-39	4/19 (21)	0.7344	5/19 (26)	0.7344	7/19 (37)	0.6300
	40-101	7/19 (37)		6/19 (32)		8/19 (42)	
	102-367	6/20 (30)		3/20 (15)		10/20 (50)	
	367-1483	7/20 (35)		5/20 (25)		6/20 (30)	

* Note: this does not include data that was not reported, and some studies fit into more than one category. P-value based on Fisher exact test.

Table 8. Subgroup analysis for retrospective datasets: Venous Thromboembolism.

		Retro res		*Retro res no KTs*		*Retro res no neg KTs*	
		found/total (%)	**P-value**	**found/total (%)**	**P-value**	**found/total (%)**	**P-value**
Study design	RCT	19/41 (46)	0.095	18/41 (44)	0.105	23/41 (56)	0.220
	NRCS	2/12 (17)		2/12 (17)		2/12 (17)	
Key Question	KQ 1	14/34 (41)	0.086	15/34 (44)	***0.007***	20/34 (59)	***0.008***
	KQ 2	1/5 (20)		1/5 (20)		2/5 (40)	
	KQ 3	8/13 (62)		2/13 (15)		4/13 (31)	
	KQ 4	2/11 (18)		1/11 (9)		2/11 (18)	
	KQ 5	7/11 (64)		8/11 (73)		9/11 (82)	
	KQ 6	0/3 (0)		0/3 (0)		0/3 (0)	
Study size	24-120	2/13 (15)	0.223	1/13 (8)	0.077	3/13 (23)	0.300
	121-716	6/14 (43)		7/14 (50)		8/14 (57)	
	717-1,532	7/13 (54)		6/13 (46)		7/13 (54)	
	1,533-316,495	6/14 (43)		6/14 (43)		7/14 (50)	

* Note: this does not include data that was not reported, and some studies fit into more than one category. P-value based on Fisher exact test.

Concept Maps and Natural Language Queries

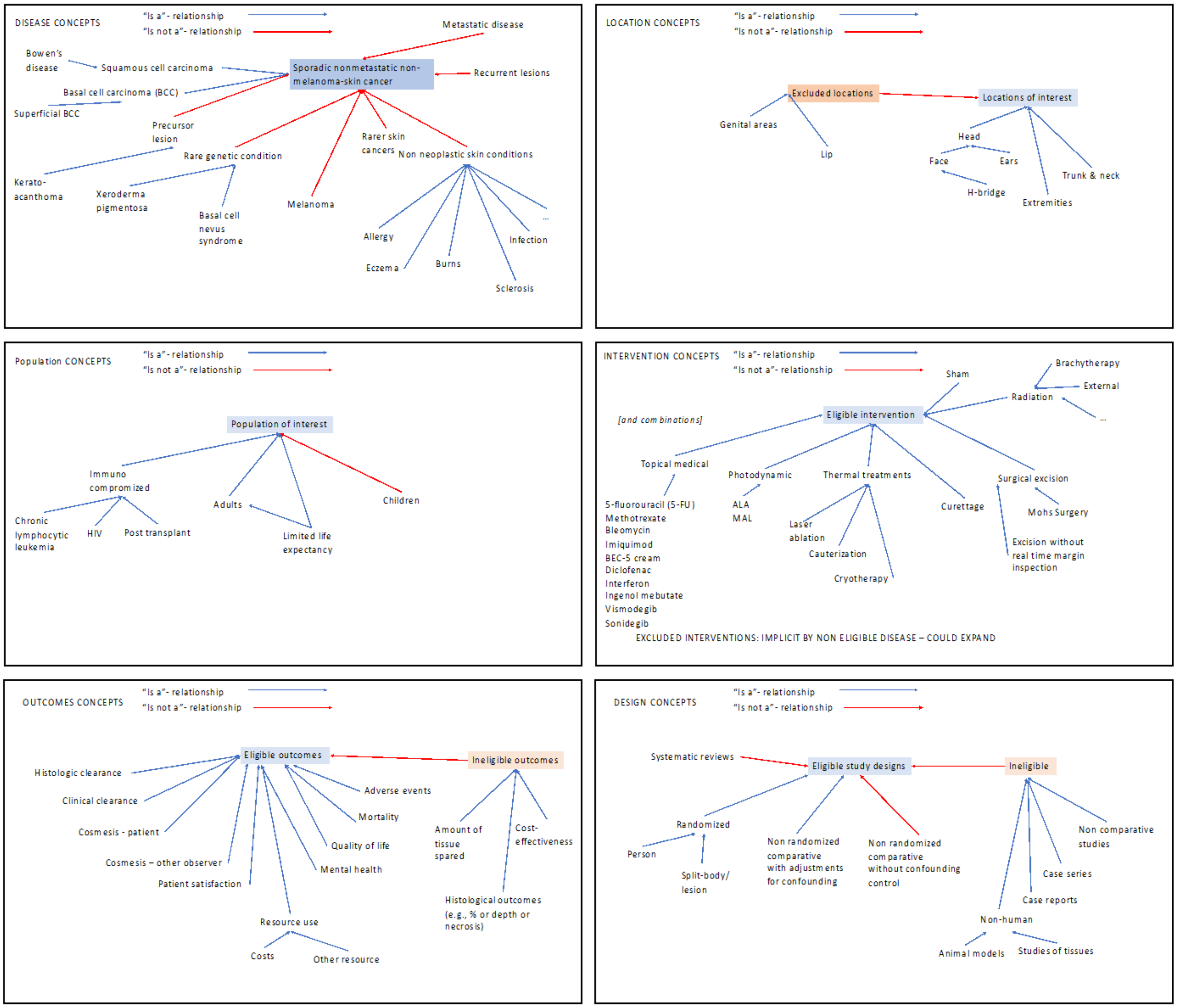
For two of the retrospective topics, we have created concept maps that represent the implicit domain knowledge that a reviewer brings to the search and screening process. In looking at these maps, it becomes clear that the natural language questions we relied on in this project are unlikely to be sufficiently descriptive to ensure adequate sensitivity. This may have accounted for the lack of sensitivity across reviews.

The Nonmelanoma Skin Cancer report had two key questions, which translated into two natural language questions:

1. For adult patients with basal cell and squamous cell carcinoma of the skin, what is the comparative effectiveness of various interventions, overall and in subgroups of interest?
2. For adult patients with basal cell and squamous cell carcinoma of the skin, how do the adverse events associated with the various interventions compare overall and in subgroups of interest?

However, as the concept map in Figure 3 shows, these seemingly clear questions are translated by humans into a complex network of concepts and terms, pertaining to various aspects of the population (disease concepts, location concepts, and population concepts), intervention, outcomes, and study design.

Figure 3. Concept Map: Nonmelanoma Skin Cancer

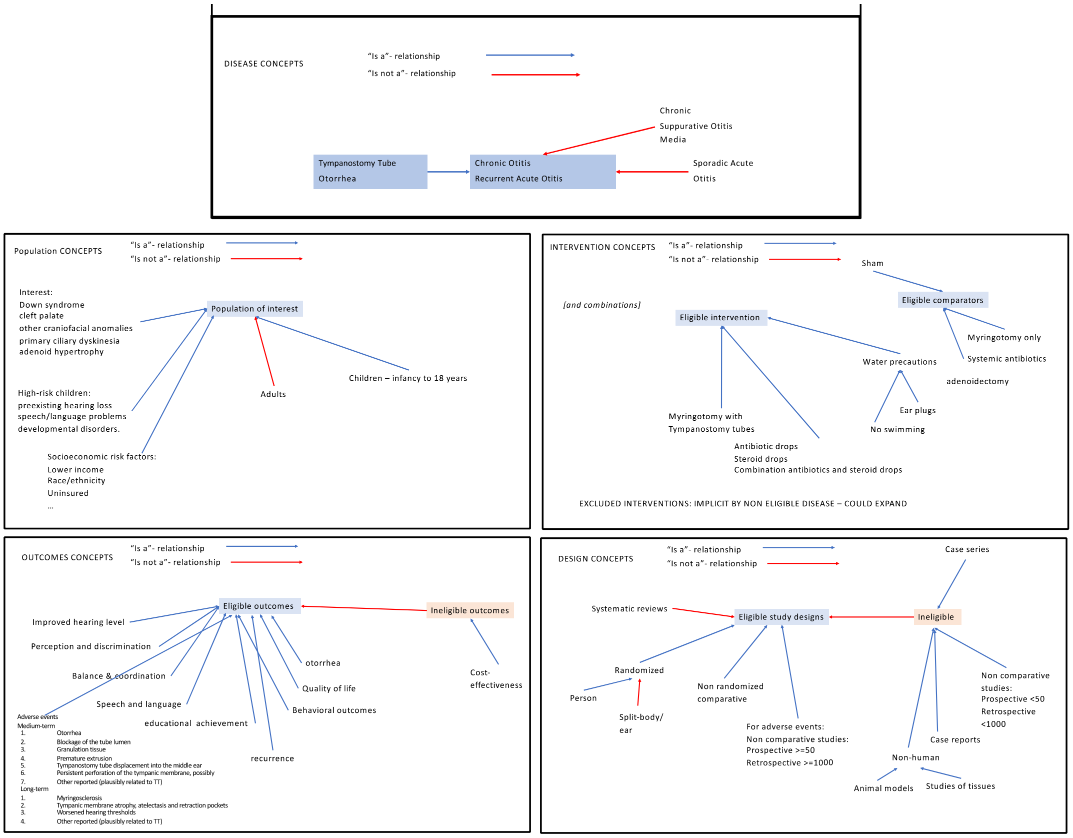


The Tympanostomy Tubes report had five key questions, which translated into eight natural language questions:

1. For children with chronic otitis media with effusion, what is the effectiveness of tympanostomy tubes, compared to watchful waiting, on resolution of middle ear effusion, hearing and vestibular outcomes, quality of life and other patient-centered outcomes?
2. What factors (such as age, age of onset, duration of effusion, comorbidities, and sociodemographic risk factors) predict which children are likely to benefit most from the intervention?
3. Does obtaining a hearing test help identify which children are more likely to benefit from the intervention?
4. For children with recurrent acute otitis media, what is the effectiveness of tympanostomy tubes, compared to watchful waiting with episodic or prophylactic antibiotic therapy, on the frequency and severity of otitis media, quality of life, and other patient centered-outcomes?
5. What factors (such as age, age of onset, number of recurrences, presence of persistent middle ear effusion, comorbidities, and sociodemographic risk factors, history of complications of acute otitis media, antibiotic allergy or intolerance) identify children who are most likely to benefit from the intervention?
6. What adverse events, surgical complications, and sequelae are associated with inserting tympanostomy tubes in children with either chronic otitis media with effusion or recurrent acute otitis media?
7. Do water precautions reduce the incidence of tympanostomy tube otorrhea, or affect quality of life?
8. In children with tympanostomy tube otorrhea, what is the comparative effectiveness of topical antibiotic drops versus systemic antibiotics or watchful waiting on duration of otorrhea, quality of life, or need for tube removal?

The concept map in Figure 4 shows the mapping of these questions onto the framework of the report protocol. This conceptual map is further complicated by the interrelationship between concepts in different key questions. For example, otorrhea is an outcome in questions 6 and 7 but part of the population in question 8.

Figure 4. Concept Map: Tympanostomy



User experience of the interface

In general, the interface we used worked well, but annotation of PICO elements for relevant abstracts added time to the screening process; one screener estimated that the screening took about twice as much time as screening the same number of citations in Abstrackr. Additionally, the process of rerunning the algorithm between batches had to be done manually, which further slowed the process. Automation of this step is key to improving the usability of the system. Some additional usability thoughts include the following:

1. Having abstracts appear multiple times (for relevant KQs) is not helpful, as the abstract will always be labeled the same way (for fear of inadvertently excluding something relevant because it is under the wrong KQ). Perhaps it would be better to list the KQs in the annotation panel, and for relevant articles the user would tag the PMID with all relevant KQs, as well as (or instead of) the tag indicating relevance.
2. It would be nice to be able to easily go back and change a decision while screening a given set; currently each abstract must be locked individually, either as it is screened or at the end of the process.
3. The text should wrap, so a reader does not have to scroll.
4. It may make sense to let the system know when you are rejecting an otherwise eligible (from a PICO standpoint) article for something like a low N or insufficient follow-up duration.

Discussion

This project sought to combine searching and screening into a single process, thereby saving time and increasing the likelihood that all citations would be identified. In testing the system prospectively and retrospectively across several completed systematic review projects, we found that while the burden was substantially decreased in most reviews, the sensitivity of the system to retrieve the relevant abstracts was unstable, ranging between 8 and 58 percent. We believe that this result is, at least in part, due to the translation and formulation of the key questions as natural language questions, where large amounts of domain knowledge are implicit. An expert can easily infer this knowledge, but machines cannot adequately encode and parse it. Future research should explore ways to encode this domain knowledge in query formulation, possibly by incorporating a "reasoning" aspect to the system to elicit more contextual information and leveraging ontologies and knowledge bases to better enrich the questions used in the search.

Limitations

The evaluation of this system was limited by the small number of prospective reviews (n=3), and the fact that labeling for each review was done by a single screener. The retrospective analyses (n=4) were also limited in that we could only provide labels for the abstracts that had also been screened by the original review team. Future evaluations should include more prospective reviews, as well as double screening and consensus adjudication of conflicts to better reflect real-world practice.

Conclusions

Both the Cochrane Handbook[13] and the AHRQ Methods Guide[14] recommend that Systematic Review searches be as comprehensive as possible within time and budget constraints. As the body of literature to be screened increases, tools that leverage machine learning have become increasingly useful in prioritizing screening based on the likelihood of relevance. The system described in this report aims to use active machine learning technology to combine the search and screening steps of systematic review, thereby improving comprehensiveness and reducing screening burden. Based on the findings of our evaluation, this technology has promise. Future work should focus on improving recall, specifically in terms of improving the ability of the algorithm to parse and contextualize natural language queries.

References

1. Elliott JH, Mavergames C, Becker L, et al. The efficient production of high quality evidence reviews is important for the public good. Bmj. 2013 Feb 13;346:f846. doi: 10.1136/bmj.f846. PMID: 23407729.

2. Tsafnat G, Dunn A, Glasziou P, et al. The automation of systematic reviews. Bmj. 2013 Jan 10;346:f139. doi: 10.1136/bmj.f139. PMID: 23305843.

3. Wallace BC, Dahabreh IJ, Schmid CH, et al. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. J Comp Eff Res. 2013 May;2(3):273-82. doi: 10.2217/cer.13.17. PMID: 24236626.

4. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010 Sep 21;7(9):e1000326. doi: 10.1371/journal.pmed.1000326. PMID: 20877712.

5. Institute of Medicine Committee on Standards for Systematic Reviews of Comparative Effectiveness R. In: Eden J, Levit L, Berg A, Morton S, eds. Finding What Works in Health Care: Standards for Systematic Reviews. Washington (DC): National Academies Press (US)

Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011.

6. Lefebvre C, Glanville J, Wieland LS, et al. Methodological developments in searching for studies for systematic reviews: past, present and future? Syst Rev. 2013 Sep 25;2:78. doi: 10.1186/2046-4053-2-78. PMID: 24066664.

7. Balk EM, Adam GP, Cao W, et al. AHRQ Comparative Effectiveness Reviews. Management of Colonic Diverticulitis. Rockville (MD): Agency for Healthcare Research and Quality (US); 2020.

8. Balk ET, T. Protocol: Continuous Positive Airway Pressure Treatment for Obstructive Sleep Apnea in Medicare Eligible Patients. 2020.

9. Saldanha IJ, Roth JL, Chen KK, et al. AHRQ Comparative Effectiveness Reviews. Management of Primary Headaches in Pregnancy. Rockville (MD): Agency for Healthcare Research and Quality (US); 2020.

10. Drucker A, Adam GP, Langberg V, et al. AHRQ Comparative Effectiveness Reviews. Treatments for Basal Cell and Squamous Cell Carcinoma of the Skin. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.

11. Steele D, Adam GP, Di M, et al. AHRQ Comparative Effectiveness Reviews. Tympanostomy Tubes in Children With Otitis Media. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.

12. Balk E, Adam GP, Kimmel H, et al. AHRQ Comparative Effectiveness Reviews. Nonsurgical Treatments for Urinary Incontinence in Women: A Systematic Review Update. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018.

13. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions version 6.0: Cochrane; 2019.

14. Relevo R, Balshem H. Finding evidence for comparing medical interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews 2011. doi: https://www.ncbi.nlm.nih.gov/books/NBK53479/pdf/Bookshelf_NBK53479.pd. PMID: 21433408.

List of publications and Products

1. Annotation tool: https://inception-project.github.io/
2. We deployed it in our platform in ATHENA RC called "Clarin": https://www.clarin.gr/en